Artificial Intelligence and Law **9:** 1–28, 2001.

Argument construction and reinstatement in logics for defeasible reasoning

JOHN F. HORTY
Philosophy Department and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, U.S.A.
E-mail: horty@umiacs.umd.edu, www.umiacs.umd.edu/users/horty

Abstract. This paper points out some problems with two recent logical systems – one due to Prakken and Sartor, the other due to Kowalski and Toni – designed for the representation of defeasible arguments in general, but with a special emphasis on legal reasoning.

1. Introduction

In recent years, researchers in nonmonotonic logic have turned increasing attention to formal systems in which nonmonotonic reasoning is analyzed through the study of interactions among competing defeasible arguments; a survey appears in Prakken and Vreewsijk (forthcoming). These *argument systems* are promising for several reasons. First, they often allow a more natural treatment of priorities among conflicting defeasible rules than the standard fixed-point or model-preference approaches, such as default logic or circumscription. Second, the explicit emphasis on the manipulation and comparison of arguments – finite syntactic entities – suggests immediate implementational possibilities; these systems are often developed within the logic programming paradigm. Finally, the formal study of relations among conflicting arguments is particularly important for the application of techniques from artificial intelligence to fields in which adversarial reasoning figures prominently, such as negotiation or, of course, the law.

I focus in this paper on two recent argument systems, both of which are heavily indebted to the work of Dung (1995) and Bondarenko et al. (1997). The first of these systems is due to Prakken and Sartor, with an initial proposal appearing in Prakken and Sartor (1996) and more elaborate developments in Prakken and Sartor (1996, 1997). This system, which will be referred to as the *PS logic*, extends the standard language of logic programming with strong negation and a connective representing defeasible implication; it allows for reasoning with prioritized defeasible rules and also for reasoning about the priorities themselves that govern these rules. The system has been provided with a fixed-point semantics, as well as a proof theory in the form of a "dialogue game", intended to model the structure

of a dispute; it was originally motivated through applications to rule-based legal reasoning, but it has also been applied to the problem of reasoning with legal precedents (Prakken and Sartor 1998).

The second system considered here is presented by Kowalski and Toni (1996) as an application of the theory set out in Bandarenko et al. (1997). This system, which will be referred to as the *KT logic*, is not yet as fully developed as that of Prakken and Sartor, but is, in principle, equally expressive, also allowing for reasoning about the priorities among defeasible rules. The system is provided in Kowalski and Toni (1996) only with a semantics, but it inherits the proof theory of Dung et al. (1996), and again, it is motivated primarily with examples involving legal reasoning.

Although the intuitions underlying many argument systems often seem to be obscured behind a cloud of syntactic complexity, the two logics studied here are clear and straightforward; they offer a real advance in our understanding of defeasible argumentation. Nevertheless, my remarks in this note are largely critical. I set out some simple examples in which it seems that these systems fail to deliver the correct results, and I explore the reasons for this failure.

The examples are drawn from the domain of defeasible inheritance networks, which has functioned as a testbed for theories of defeasible reasoning ever since the initial exploration of Criscuolo, Etherington, and Reiter in papers such as Etherington (1987), Etherington and Reiter (1983), and and Reiter and Criscuolo (1981). Of course, the two logics studied here are formulated in a much richer language than that of defeasible inheritance networks. Still, even though they are substantially more expressive, these logics should yield correct results when restricted to the simple language of inheritance; and this, I argue, they fail to do. In particular, I show that the results of these logic conflict with the account of inheritance reasoning developed by Thomason, Touretzky, and myself in a series of papers including Horty (1994), Horty and Thomason (1988), Horty et al. (1990), and Touretzky et al. (1987, 1991).

2. The PS Logic

The language of the PS logic extends that of standard logic programming by allowing strong (or classical) negation in addition to the usual weak negation (or negation by failure), and by allowing rules to be formed through defeasible as well as strict implication. The logic is given two formulations: in the first, the priorities among defeasible rules are specified as part of the background theory, through a fixed ordering; in the second, more general formulation, these priorities might themselves be established through defeasible reasoning.

Although, as Prakken and Sartor show, this expressive richness is useful for representing realistic legal argument, the background language can be simplified considerably once we restrict our attention to simple inheritance reasoning: there is then no need for weak negation, no need, even, for rules containing multiple

antecedents; and we can make do with the first formulation of the logic, with fixed priorities among defeasible rules. As a result, the presentation of the system can be simplified considerably.

In order to isolate the aspects of the PS logic that will concern us here, I set out in this section a simple and streamlined version of the system, adequate only for the formalization of inheritance reasoning. Given the linguistic restrictions, the reader can verify that the theory described here is a special case of the full logic.

2.1. LITERALS, RULES, AND THEORIES

We begin with a description of the simple language.

A *literal* is either an atomic formula A or a formula $\neg A$ with A atomic. We say that A and $\neg A$ are *complements*, and where L is a literal, that $\overline{L}$ is the complement of L. A special statement $\top$ is singled out to represent truth.

Literals can be combined through either strict or defeasible implication connectives to form rules. A *strict rule* has the form

$$r_i : \quad L_j \Rightarrow L_k,$$

while a *defeasible rule* has the form

$$r_i : \quad L_j \rightarrow L_k,$$

where, in each case, r_i is a label of the rule (and is often omitted in presentation).[1] An assertion is represented by a rule having truth as its antecedent: the categorical assertion of the literal L, for example, is carried by the strict rule $\top \rightarrow L$, while the assertion that L holds by default is carried by the defeasible rule $\top \rightarrow L$. In order to avoid complications having to do with unification, we follow the familiar practice of avoiding variables in rules, expressing a general statement through the collection of its instances.

An *ordered theory* is a tuple of the form $\Gamma = \langle S, D, < \rangle$, where S is a set of strict rules, D a set of defeasible rules, and $<$ a partial ordering representing priority on the defeasible rules: if $r_i < r_j$, the defeasible rule r_j is taken to have a higher priority than r_i, and should be preferred in case of conflict.

In order to illustrate the way in which inheritance networks can be interpreted as ordered theories, we consider the ordered theories corresponding to two familiar networks, the Tweety Triangle and the Nixon Diamond. The first is depicted in Figure 1, with Pt, Bt, and Ft representing the propositions that Tweety is a penguin, a bird, and a flying thing; more exactly, this first theory is the tuple $\langle S, D, < \rangle$ with

[1] This use of $\Rightarrow$ and $\rightarrow$ to represent strict and defeasible implication respectively follows the convention established in Horty and Thomason (1988) for inheritance networks, and seems to make typographical sense as well, with the visually stronger arrow representing the logically stronger form of implication. Prakken and Sartor adopt the opposite convention, using $\Rightarrow$ to represent defeasible and $\rightarrow$ strict implication.

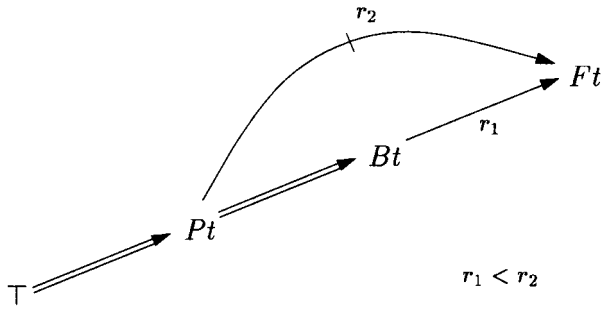

Figure 1. The Tweety Triangle.

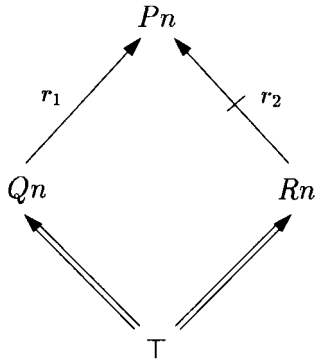

Figure 2. The Nixon Diamond.

$S = \{\top \rightarrow Pt, Pt \rightarrow Bt\}$, with $D = \{r_1 : Bt \rightarrow Ft, \quad r_2 : Pt \rightarrow \neg Ft, \}$, and with $r_1 < r_2$.[2] The second of these two theories is depicted in Figure 2, with Qn, Rn, and Pn representing the propositions that Nixon is a Quaker, a Republican, and a pacifist; more exactly, this theory is the tuple $\langle S, D, < \rangle$ with $S = \{\top \Rightarrow Qn, \top \Rightarrow Rn\}$, with $D = \{r_1 : Qn \rightarrow Pn, \quad r_2 : Rn \rightarrow \neg Pn\}$, and with no priority relations holding among the two defeasible rules.

2.2. ARGUMENTS, CONFLICT, AND DEFEAT

We now turn to the arguments that can be constructed on the basis of the background language. Because we have simplified the language so thoroughly – and in particular, because we have restricted the rules to contain only single literals as antecedents – we can make due with a linear notion of argument, as follows.

- An *argument* based on an ordered theory $\Gamma = \langle S, D, < \rangle$ is a finite sequence $\alpha = [r_0, \ldots, r_n]$ of rules from $S \cup D$ such that: (i) the antecedent of r_1 is $\top$; (ii)

[2] In our graphical depiction of ordered theories, rules of the form $L_j \rightarrow \neg L_k$ and $L_j \Rightarrow \neg L_k$ are drawn in a way that conforms to the standard inheritance notation, as $L_j \not\rightarrow L_k$ and $L_j \not\Rightarrow L_k$ respectively.

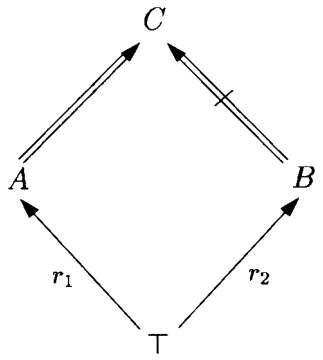

Figure 3. An implicit conflict.

the antecedent of r_{i+1} is identical with the consequent of r_i for all $0 \geq i \geq n$; and (iii) no two rules from α have the same consequent.

The first two conditions of this definition guarantee that an argument based on a theory begins with an assertion immediately supported by that theory, and then proceeds through instances of strict and defeasible modus ponens; the third condition is intended simply to rule out incidental complications involved in the consideration of cyclic arguments. We illustrate the definition by noting that the sequence

$$\alpha = [\top \Rightarrow Pt, \quad Pt \Rightarrow Bt, \quad r_1 : Bt \rightarrow Ft]$$

is an argument based on the Tweety Triangle from Figure 1. Since the arguments that concern us here are linear, and in order to further emphasize the relations with inheritance theory, we will depict these arguments in a manner analogous to the familiar path notation from the inheritance literature; thus, for example, the argument α above will be depicted as the path

$$\top \Rightarrow Pt \Rightarrow Bt \underset{r_1}{\rightarrow} Ft.$$

Where Γ is an ordered theory, we let $Args_\Gamma$ represent the set of arguments based on that theory, the set of arguments that can be constructed using the materials contained in Γ. Like rules, arguments themselves can be classified as *strict* or *defeasible* – strict if they contain no defeasible rules, and defeasible if they do. Finally, the set of *conclusions* of an argument are defined as the set of literals contained in that argument, so that, for example, the argument α above has as its conclusions the literals: $\top$, Pt, Bt, and Ft.

Arguments based on ordered theories, such as those depicted in the Tweety and Nixon examples, can conflict with one another; and it is at first tempting to think of two arguments as conflicting just in case they contain complementary literals as conclusions. This notion of conflict fails to account, however, for those implicit conflicts that might be induced by the strict rules contained in an ordered theory.

Consider Figure 3, for example. Here the two arguments $\top \underset{r_1}{\rightarrow} A$ and $\top \underset{r_2}{\rightarrow} B$ do not themselves contain complementary literals, but it is all the same natural to think of them as conflicting. Anyone who accepts the first of these arguments is committed to accepting its extension $\top \underset{r_1}{\rightarrow} A \Rightarrow C$, anyone who accepts the second is likewise committed to the extension $\top \underset{r_2}{\rightarrow} B \Rightarrow \neg C$, and these two extended arguments do stand in explicit conflict.

Motivated by examples like this, Prakken and Sartor define a notion of conflict among arguments in a way that takes their strict extensions into account; in the current, simplified setting, the idea behind their definition can be presented as follows. First, where α is an argument and σ is a sequence of strict rules, we let $\alpha + \sigma$ represent the concatenation of α and σ. Then, where α and α' are two arguments based on an ordered theory Γ, we can say that α *conflicts with* α' just in case there are strict sequences σ and σ' and complementary literals L and $\overline{L}$ such that: $\alpha + \sigma$ is an argument based on Γ with conclusion L, and $\alpha' + \sigma'$ is an argument based on Γ with conclusion $\overline{L}$.

In fact, we are not interested so much in conflict as in the related notion of defeat, where, intuitively, one argument defeats another if it conflicts with and is not weaker than that argument. Again, because of their rich background language, the notion of defeat set out by Prakken and Sartor is complicated, relying upon a number of preliminary technical concepts (such as rebutting and undercutting), but in the current setting we can take a more direct route.

Concerning argument strength, we assume, following Prakken and Sartor, that all strict arguments are equally strong, and stronger than all defeasible arguments, and that the strength of a defeasible argument in support of a particular conclusion is measured by the strength of the final defeasible rule supporting that conclusion. In order to capture this idea precisely, let us first introduce a special symbol ∞ representing infinite strength, the strength of a strict argument, and so placed in the priority ranking above any defeasible rule (that is, $r < \infty$ for each defeasible rule r). Adapting Prakken and Sartor's notation, we suppose, where α is an argument with conclusion L, that $R_L(\alpha)$ is defined as follows: if the subargument of α prior to the occurrence of L is strict, then $R_L(\alpha) = \infty$; otherwise, $R_L(\alpha)$ refers to the final defeasible rule in α that either contains L as consequent or occurs entirely prior to L. (Example: if $\alpha = \top \Rightarrow A \underset{r_1}{\rightarrow} B \underset{r_2}{\rightarrow} C \Rightarrow D \underset{r_3}{\rightarrow} E$, we have $R_E(\alpha) = r_3$, $R_D(\alpha) = r_2$, and $R_A(\alpha) = \infty$.) The strength of the argument α in support of the conclusion L can then be measured by $R_L(\alpha)$.

With this notion of strength in place, we can now specialize Prakken and Sartor's notion of defeat to our simple setting.

- Let α and α' be two arguments based on an ordered theory Γ. Then α *defeats* α' just in case there are strict sequences σ and σ' and complementary literals L and $\overline{L}$ such that: (i) $\alpha + \sigma$ is an argument based on Γ with conclusion L, (ii) $\alpha' + \sigma'$ is an argument based on Γ with conclusion $\overline{L}$, and (iii) it is not the case that $R_L(\alpha + \sigma) < R_{\overline{L}}(\alpha' + \sigma')$.

The first two clauses of this definition, which are carried over from the definition of conflict, tell us simply that the arguments α and α' conflict with one another concerning the conclusions L and $\overline{L}$; the third clause tells us that the support provided by α for L is not weaker than that provided by α' for $\overline{L}$.

It is possible, of course, for each of two arguments to defeat the other. This occurs, for example, in the Nixon Diamond, where each of the two arguments $\top \Rightarrow Qn \underset{r_1}{\rightarrow} Pn$ and $\top \Rightarrow Rn \underset{r_2}{\rightarrow} \neg Pn$ defeats the other, since they support the complementary literals Pn and $\neg Pn$, yet neither offers stronger support than the other; we have neither $r_1 < r_2$ nor $r_2 < r_1$. Unlike the relation of conflict, however, the defeat relation is not necessarily symmetric. In the case of the Tweety Triangle, for instance, the argument $\top \Rightarrow Pt \underset{r_2}{\rightarrow} \neg Ft$ defeats the argument $\top \Rightarrow Pt \Rightarrow Bt \underset{r_1}{\rightarrow} Ft$, since these two arguments support the complementary literals $\neg Ft$ and Ft and the support provided by the second is not stronger than that provided by the first; we do not have $r_2 < r_1$. On the other hand, the second of these two arguments does not defeat the first, since the support provided by the first is stronger than the support provided by the second; here, we do have $r_1 < r_2$.

In a case like this, where one argument α defeats another argument α', but α' does not defeat α, we say that α *strictly defeats* α'.

2.3. ACCEPTABILITY AND JUSTIFICATION

The general picture of defeasible reasoning underlying the PS logic is this. Given an ordered theory Γ, one first constructs the entire set $Args_\Gamma$ of arguments based on Γ, and then computes the defeat relations among these arguments. On the basis of this pattern of defeat relations, one then isolates a particular subset of the arguments that are to count as justified. The conclusions supported by the theory are the conclusions of these justified arguments.

In our simple setting, we now know how to construct the set of arguments $Args_\Gamma$ from the theory Γ, and also how to define the defeat relations among the members of $Args_\Gamma$; and of course, it is a simple matter, once the set of justified arguments is isolated, to collect their conclusions. All that is missing, therefore, is a definition that tells us, given the pattern of defeat relations among arguments, which of these arguments are to be regarded as justified.

It may seem tempting to suppose that an argument should be regarded as justified if it is, in fact, not defeated. Prakken and Sartor, however, base their theory on the more complicated idea of *reinstatement* – deriving most immediately from Dung (1995), but going back at least to Pollock (1987) – according to which even certain defeated arguments should be regarded as justified, as long as the arguments defeating them are themselves defeated. Suppose, for example, that a theory allows for the construction of three arguments α_1, α_2, and α_3, subject to the following defeat relations: α_2 defeats α_1, and α_3 defeats α_2. In such a case, according to Prakken and Sartor, the argument α_3 should be thought of as reinstating α_1, by

defeating the only argument that defeats it, so that α_1 itself is to be regarded as justified.

The particular way in which reinstatement enters into the PS logic is through the concept of acceptability, closely related to that of Dung, and defined here as follows.

- Let Γ be an ordered theory and $\mathcal{S}$ a subset of $Args_\Gamma$. Then an argument α is *acceptable* with respect to $\mathcal{S}$ just in case each argument α' that defeats α is itself strictly defeated by some argument α'' belonging to $\mathcal{S}$.

The idea, of course, is that α should be acceptable with respect to $\mathcal{S}$ whenever α either is not defeated at all or, if defeated, is reinstated by some argument already belonging to $\mathcal{S}$.

Given this notion of acceptability, Prakken and Sartor then define a function, again due to Dung, that maps each set of arguments into the set containing those arguments that are acceptable with respect to the initial set.

- The *characteristic function* of an ordered theory Γ is the function F_Γ where, for each subset $\mathcal{S}$ of $Args_\Gamma$,

$$F_\Gamma(\mathcal{S}) = \{\alpha \in Args_\Gamma : \alpha \quad \text{is} \quad \text{acceptable} \quad \text{with} \quad \text{respect} \quad \text{to} \quad \mathcal{S}\}.$$

It is easy to see that the function F_Γ is monotonic on subsets of $Args_\Gamma$, under the subset relation. By the Knaster–Tarski theorem, it therefore has a least fixed point; and Prakken and Sartor suggest that the arguments that are justified on the basis of Γ – represented as $JustArgs_\Gamma$ – should be defined as the members of this least fixed point.

- Where Γ is an ordered theory, the set $JustArgs_\Gamma$ is the least fixed point of the characteristic function F_Γ.

Prakken and Sartor discuss various desirable properties of the justified arguments defined in this way, and they note, following Dung, that in the finitary case, where each argument is defeated by at most a finite number of arguments, the set of justified arguments can be defined through an iterative construction. More exactly, where Γ is an ordered theory and the sequence $F^1_\Gamma, F^2_\Gamma, F^3_\Gamma, \ldots$ is defined by taking

$$\begin{aligned} F^1_\Gamma &= F_\Gamma(\emptyset), \\ F^{i+1}_\Gamma &= F^i_\Gamma \cup F_\Gamma(F^i_\Gamma), \end{aligned}$$

it can be shown that $\bigcup_{i=1}^{\infty}(F^i_\Gamma) = JustArgs_\Gamma$.

We illustrate the various definitions underlying the PS logic by applying this iterative construction to the Tweety and Nixon examples. Beginning with Tweety, let Γ represent the ordered theory depicted in Figure 1. Then $Args_\Gamma$ contains the arguments

$$\begin{aligned} \alpha_1 &= \top \Rightarrow Pt, \\ \alpha_2 &= \top \Rightarrow Pt \Rightarrow Bt, \\ \alpha_3 &= \top \Rightarrow Pt \Rightarrow Bt \underset{r_1}{\rightarrow} Ft, \\ \alpha_4 &= \top \Rightarrow Pt \underset{r_2}{\rightarrow} \neg Ft, \end{aligned}$$

with the only defeat relation being: α_4 strictly defeats α_3. Because α_1, α_2, and α_4 are undefeated, it is easy to see that $F_\Gamma^1 = \{\alpha_1, \alpha_2, \alpha_4\}$, and that $F_\Gamma^2 = F_\Gamma^1$. This set, therefore, is the least fixed point of the function F_Γ, containing the justified arguments. Because it classifies α_4 as justified, the theory thus yields the intuitively correct result that Tweety flies.

Turning to Nixon, we let Γ represent the ordered theory depicted in Figure 2, so that $Args_\Gamma$ contains the arguments

$$\begin{aligned}
\alpha_1 &= \top \Rightarrow Qn,\\
\alpha_2 &= \top \Rightarrow Rn,\\
\alpha_3 &= \top \Rightarrow Qn \underset{r_1}{\rightarrow} Pn,\\
\alpha_4 &= \top \Rightarrow Rn \underset{r_2}{\rightarrow} \neg Pn,
\end{aligned}$$

subject to the defeat relations: α_3 defeats α_4 and α_4 defeats α_3. It can then be seen that $F_\Gamma^1 = \{\alpha_1, \alpha_2\}$ and that $F_\Gamma^2 = F_\Gamma^1$. This set, then, is the least fixed point of F_Γ, and since it contains neither α_3, nor α_4, the theory again yields the result that is correct from a skeptical point of view, that it is unreasonable to conclude that Nixon is a pacifist but also unreasonable to conclude that he is not.

3. Problems with the PS Logic

Having considered some cases in which Prakken and Sartor's logic seems to work well, I now wish to point out some problems with system. Ignoring various matters of detail, I concentrate on what I take to be two major issues: first, a difficulty in the process of constructing and evaluating arguments; second, a difficulty with the notion of reinstatement.

I illustrate these difficulties by exhibiting some examples in which the logic seems to yield incorrect results. In each case, I argue that the results of the logic are incorrect using the only method that I know of in this area, where there is no recourse to anything like a formal semantics: telling a story, encoding the story in the logic, calculating the supported conclusions, and then relying on the reader's intuitions to coincide with my own view that the conclusions generated by the logic do not agree with those that would reasonably be drawn from the story.

3.1. ARGUMENT CONSTRUCTION AND EVALUATION

Here is the first story. Let us suppose, as may even be the case, that lawyers tend to be wealthy, but that a certain subclass of lawyers, public defenders, tend not to be wealthy. And imagine that there is an area of town – say, Brentwood – containing a large number of expensive private homes along with a much smaller number of middle-income rental properties, so that the residents of Brentwood tend to be wealthy, although a certain subclass of Brentwood residents, the renters, tend not

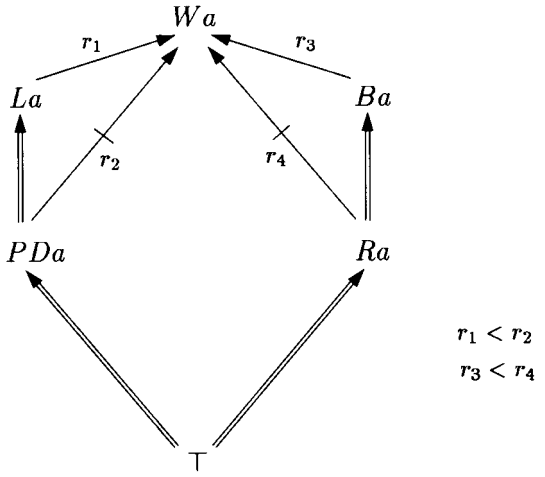

Figure 4. Is Ann wealthy?

to be. We will suppose that Ann is a public defender, and therefore a lawyer, who rents in Brentwood, and is therefore a resident of Brentwood.

The information from this story is presented in the ordered theory Γ, depicted in Figure 4, where PDa, La, Ra, Ba, and Wa represent the respective propositions that Ann is a public defender, a lawyer, a renter in Brentwood, a resident of Brentwood, and wealthy. Since we prefer information based on subclasses to information based on superclasses, we have $r_1 < r_2$ and $r_3 < r_4$ – that is, the rule that Ann is not wealthy because she is a public defender has a higher priority than the rule that she is wealthy because she is a lawyer, and the rule that Ann is not wealthy because she is a rents in Brentwood has a higher priority than the rule that she is wealthy because she is a Brentwood resident. The rules r_1 and r_4, however, are incomparable, as are the rules r_2 and r_3.

Given Γ, it is easy to see that $Args_\Gamma$ contains the following arguments:

$$\begin{aligned}
\alpha_1 &= \top \Rightarrow PDa,\\
\alpha_2 &= \top \Rightarrow Ra,\\
\alpha_3 &= \top \Rightarrow PDa \Rightarrow La,\\
\alpha_4 &= \top \Rightarrow Ra \Rightarrow Ba,\\
\alpha_5 &= \top \Rightarrow PDa \Rightarrow La \underset{r_1}{\rightarrow} Wa,\\
\alpha_6 &= \top \Rightarrow PDa \underset{r_2}{\rightarrow} \neg Wa,\\
\alpha_7 &= \top \Rightarrow Ra \Rightarrow Ba \underset{r_3}{\rightarrow} Wa,\\
\alpha_8 &= \top \Rightarrow Ra \underset{r_4}{\rightarrow} \neg Wa.
\end{aligned}$$

What about defeat relations? Well, the argument α_6 defeats α_5, since these two arguments support the conflicting literals $\neg Wa$ and Wa, and the support provided by the second is not stronger than that provided by the first: we do not have $r_2 < r_1$.

And we know likewise that α_8 defeats α_7. But it also turns out that α_5 defeats α_8 and that α_7 defeats α_6, since, again, these arguments support the conflicting literals Wa and $\neg Wa$, and the support provided by the defeated arguments is no stronger than the support provided by their defeaters: we have neither $r_1 < r_4$ nor $r_3 < r_2$. Given these defeat relations, it is easy to see that $F_\Gamma^1 = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, and then that $F_\Gamma^2 = F_\Gamma^1$, so that this set is the least fixed point of F_Γ. Neither α_6 nor α_8 is acceptable relative to this set, since each is defeated by some argument – α_7 and α_5 respectively – that is not strictly defeated by an argument already belonging to the set. The PS logic therefore does not allow us to draw any conclusions about Ann's wealth.

This result seems to run contrary to intuition, or at least, contrary to the intuitions underlying the theory of skeptical inheritance presented in Horty et al. (1990). According to this theory, neither of the arguments α_5 nor α_7 should be given any weight at all, either in supporting conclusions or in interfering with other arguments, since each of these arguments is itself strictly defeated. Therefore, the fact that these arguments defeat α_8 and α_6 should have no bearing on the acceptability of the latter; α_8 and α_6 should thus be accepted, leading to the conclusion that Ann is not wealthy.

In fact, Prakken and Sartor are sympathetic with this perspective, and suggest in their treatment of the somewhat similar Example 3.16 from their (1997), and also in personal conversation that the appropriate results might be achievable within the general framework of the PS logic with only some minor modifications. In particular, they note that the set of justified arguments supported by skeptical inheritance – that is, the set $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_7\}$ – is in fact a fixed point of the function F_Γ, but simply not the least fixed point.

My own diagnosis is that the problem is more serious. I believe it is symptom of a general strategic error involved in the design of the PS logic: the separation of the process of argument construction from that of argument evaluation. As we have seen, the PS logic reflects a picture of defeasible reasoning according to which all of the possible arguments based on a particular theory are first constructed, relations of defeat among these arguments are noted, and then the justified arguments are defined on the basis of this overall pattern of defeat. As a result, it is possible for arguments that are clearly indefensible to perturb the pattern of defeat so as to affect arguments that, from an intuitive point of view, should count as justified. The alternative strategy – followed in the inheritance literature – involves interleaving the construction and evaluation of arguments. Arguments are constructed step by step, and are evaluated after each step in their construction; those that are indefensible, such as α_5 and α_7 above, are discarded at once, and so cannot influence the status of others.

A more striking example of the problems resulting from the separation of argument construction and evaluation is presented in Figure 5, which can be given an intuitive interpretation by taking *RCb*, *RNb*, *CCb*, *CUb*, and *VUb* to represent the respective propositions that Bob is a resident of Cuba, a resident of North America,

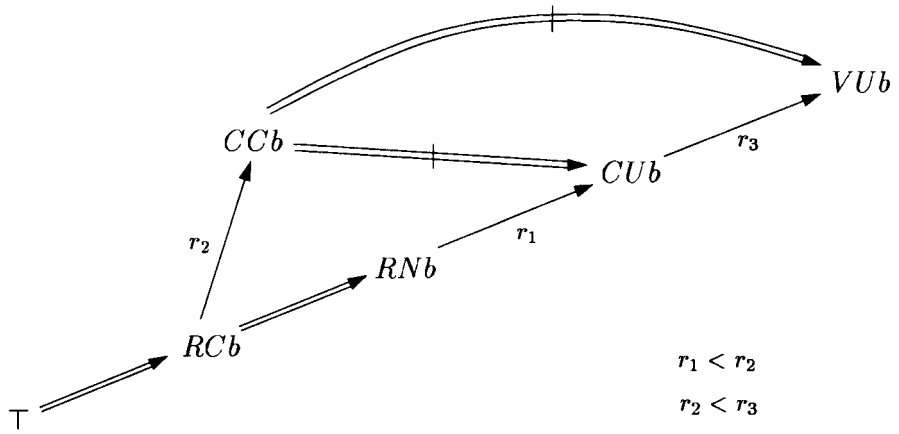

Figure 5. Can Bob vote in the US?

a citizen of Cuba, a citizen of the US, and a person with voting rights in the US. The story that goes along with this diagram is as follows. First, we are to assume that Bob is a resident of Cuba. Second, that residents of Cuba are residents of North America; this holds by definition, since Cuba is a part of North America. Third, there is a weak default – with some statistical justification – according to which residents of North America tend to be citizens of the US. Fourth, there is a stronger default according to which residents of Cuba tend to be citizens of Cuba. Fifth and sixth, citizens of Cuba cannot also be citizens of the US, or have voting rights in the US. Seventh, there is a very strong default – stronger than any of the others, and violated, as far as I know, only by convicted felons – according to which citizens of the US tend to have voting rights in the US.

Given this story, what does the PS logic tell us about Bob? If the depicted theory is Γ, then $Args_\Gamma$ contains the following arguments:

$$
\begin{aligned}
\alpha_1 &= \top \Rightarrow RCb,\\
\alpha_2 &= \top \Rightarrow RCb \Rightarrow RNb,\\
\alpha_3 &= \top \Rightarrow RCb \Rightarrow RNb \underset{r_1}{\rightarrow} CUb,\\
\alpha_4 &= \top \Rightarrow RCb \Rightarrow RNb \underset{r_1}{\rightarrow} CUb \underset{r_3}{\rightarrow} VUb,\\
\alpha_5 &= \top \Rightarrow RCb \underset{r_2}{\rightarrow} CCb,\\
\alpha_6 &= \top \Rightarrow RCb \underset{r_2}{\rightarrow} CCb \Rightarrow \neg CUb,\\
\alpha_7 &= \top \Rightarrow RCb \underset{r_2}{\rightarrow} CCb \Rightarrow \neg VUb.
\end{aligned}
$$

Turning to defeat relations, let us consider for simplicity only the three arguments ending with defeasible inferences – α_3, α_4, and α_5 – since the defeat relations among the other defeasible arguments, α_6 and α_7, are determined by these. It is clear, first, that α_5 strictly defeats α_3. For the argument α_6, which is simply a strict extension of α_5, supports the conclusion $\neg CUb$, which conflicts with the

conclusion *CUb* supported by α_3; and since $R_{CUb}(\alpha_3) = r_1$ and $R_{\neg CUb}(\alpha_6) = r_2$, the default supporting $\neg CUb$ in α_6 is stronger than the default supporting *CUb* in α_3. For the same reason, it is clear that α_5 defeats α_4, because α_4 likewise supports *CUb* through a default rule weaker than that through which α_7 supports $\neg CUb$. In this case, however, the defeat is not strict, for it turns out that α_4 also defeats α_5. Why? Well, the argument α_7, which is again a strict extension of α_5, supports the conclusion $\neg VUb$, which conflicts with the conclusion *VUb*; and since $R_{\neg VUb}(\alpha_7) = r_2$ and $R_{VUb}(\alpha_4) = r_3$, the default supporting *VUb* in α_3 is stronger than that supporting $\neg VUb$ in α_7.

Given these defeat relations, the iterative procedure tells us that $F_\Gamma^1 = \{\alpha_1, \alpha_2\}$, and then that $F_\Gamma^2 = F_\Gamma^1$, so that this set is the least fixed point of F_Γ. In particular, then, the PS logic does not count the arguments α_5, α_6, and α_7 as justified, and so does not support the conclusions that Bob is citizen of Cuba, that Bob is not a citizen of the US, or that Bob does not have voting rights in the US.

Again, this result is contrary to the intuitions underlying the skeptical inheritance theory of Horty et al. (1990), which would classify all of the arguments from this example except α_3 and α_4 as justified, and therefore tell us that Bob is a citizen of Cuba, not a citizen of the US, and that he does not have voting rights in the US. Again, the reason for this disparity seems to be the separation in the PS logic between the processes of argument construction and evaluation. In this case, as we have seen, α_5 strictly defeats α_3, but then α_5 is itself defeated by α_4. What makes the example so striking is that α_4 is itself an extension of α_3, so that α_5 is defeated by an extension of an argument that it strictly defeats. In a theory that evaluates arguments immediately upon construction, α_3 would be classified as indefensible as soon as it was discovered to be strictly defeated by α_5; this argument would therefore not be available to serve as a basis for the construction of further arguments, such as α_4, which might then interfere with α_5.

Note also that Prakken and Sartor's tentative suggestion that the appropriate set of arguments might be identifiable as some fixed point of the F_Γ function, even if not the least fixed point, does not work in this case. Here, the set of justified arguments according to the skeptical theory is $\mathcal{S} = \{\alpha_1, \alpha_2, \alpha_5, \alpha_6, \alpha_7\}$, but this set is not a fixed point of F_Γ, since the argument α_5 belongs to $\mathcal{S}$ but not to $F_\Gamma(\mathcal{S})$.

3.2. REINSTATEMENT

The second difficulty I want to discuss concerns the notion of reinstatement – the idea that an argument should count as acceptable even if it is defeated, as long as all the arguments defeating it are themselves strictly defeated.

As we have seen, this idea plays an important role in the architecture of the PS logic. Its effect can be illustrated through the ordered theory displayed in Figure 6, where *WCa*, *Ca*, *Ba*, and *Fa* represent the respective propositions that Al is a wild chicken, a chicken, a bird, and a flying thing. The story that goes along with this picture is mostly familiar – as a rule, birds tend to fly and chickens tend not to –

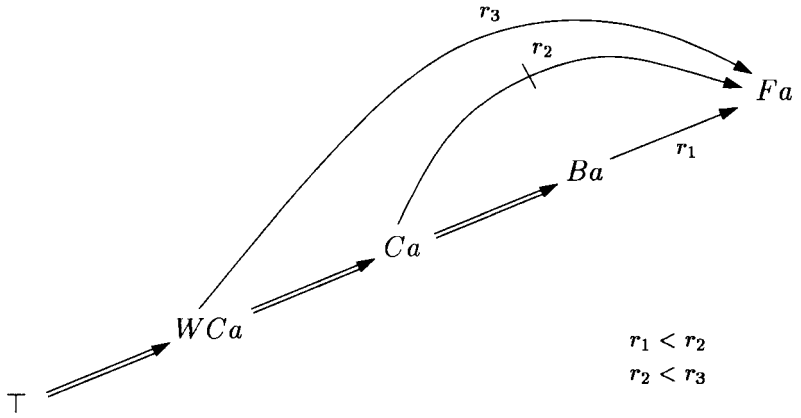

Figure 6. Does Al fly because he is a bird?

but there is one fresh twist: we have discovered a new kind of chicken, known as a wild chicken, which is able to fly, and we are told that Al is a wild chicken.[3]

Taking Γ as the depicted theory, the set $Args_\Gamma$ contains

$$\begin{aligned}
\alpha_1 &= \top \Rightarrow WCa,\\
\alpha_2 &= \top \Rightarrow WCa \Rightarrow Ca,\\
\alpha_3 &= \top \Rightarrow WCa \Rightarrow Ca \Rightarrow Ba,\\
\alpha_4 &= \top \Rightarrow WCa \Rightarrow Ca \Rightarrow Ba \underset{r_1}{\rightarrow} Fa,\\
\alpha_5 &= \top \Rightarrow WCa \Rightarrow Ca, \underset{r_2}{\rightarrow} \neg Fa,\\
\alpha_6 &= \top \Rightarrow WCa \underset{r_3}{\rightarrow} Fa.
\end{aligned}$$

Given the ordering on defaults, the defeat relations among these arguments are clear: α_5 strictly defeats α_4, and α_6 strictly defeats α_5. Even though α_4 is defeated, then, the only argument that defeats it is itself strictly defeated. The argument α_4 is therefore reinstated, and we should expect it to count as justified according to the PS logic. The iterative procedure confirms this expectation, telling us that $F^1_\Gamma = \{\alpha_1, \alpha_2, \alpha_3, \alpha_6\}$, that $F^2_\Gamma = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_6\}$, and that $F^3_\Gamma = F^2_\Gamma$, so that this set is the minimal fixed point of the F_Γ operator, containing the justified arguments.

The classification of α_4 as justified may seem odd, however – not so much because of its conclusion that Al flies, which is legitimately supported by α_6, but because the argument itself is problematic, suggesting that Al flies because he is a bird, that this is a good reason for believing him to fly. In the context, this argument appears to be flawed, since Al is a special kind of bird, a chicken, that does not fly. The reason Al flies is that he is a wild chicken, and one might wonder why, just because this reason is sufficient to justify the conclusion, another reason that has already been discredited should again be endorsed.

[3] Wild chickens were first introduced into the inheritance literature in Touretzky et al. (1991).

Still, even if one agrees with this criticism, it might seem like a minor point. Perhaps all we should worry about in evaluating an argument system is the set of conclusions it supports. Perhaps it is best not to take the arguments themselves too seriously, beyond their technical role in generating the appropriate set of conclusions; if reinstatement provides a technically convenient specification of the appropriate conclusion set, perhaps this is a sufficient justification for the principle.

Indeed, as discussed in Touretzky et al. (1991), the initial version of skeptical inheritance set out in Horty et al. (1990) did accept reinstatement primarily for reasons of technical convenience – in order to allow for the parallel marker propagation inference algorithm described in that paper. It was soon realized, however, that reinstatement had a more serious semantic impact, and could not be accepted no matter how convenient: in addition to forcing us to accept certain problematic arguments for correct conclusions, it leads to conclusions that are simply incorrect. Because of this, the skeptical inheritance theory originally presented in Horty et al. (1990) was reformulated in Horty (1994) in a way that avoids any reliance on reinstatement.

The kind of incorrect conclusion supported by reinstatement is illustrated by the following story. Imagine that, in virtue of stock options accrued over the years, most Microsoft employees are by now millionaires; imagine it is at least a weak default that Microsoft employees are millionaires.[4] Suppose also, as a slightly stronger default, that most new Microsoft employees, many of them just out of college, have not yet accumulated so much as half a million dollars. Finally, imagine that Beth is a new Microsoft employee, but suppose there is reason to believe, as a very strong default – perhaps someone has actually seen a recent list of assets – that Beth does happen to have half a million dollars. And let us supplement this defeasible information by explicitly noting the strict truths that any new Microsoft employee is necessarily a Microsoft employee, and that anyone with a million dollars also has half a million dollars.

The information from this story can be represented as the ordered theory Γ shown in Figure 7, with *NMEb*, *MEb*, *1Mb*, $^1/_2Mb$ representing the respective propositions that Beth is a new Microsoft employee, that she is a Microsoft employee, that she has a million dollars, and that she has half a million dollars. The arguments belonging to $Args_\Gamma$ are

$$
\begin{aligned}
\alpha_1 &= \top \Rightarrow \mathit{NMEb},\\
\alpha_2 &= \top \Rightarrow \mathit{NMEb} \Rightarrow \mathit{MEb},\\
\alpha_3 &= \top \Rightarrow \mathit{NMEb} \Rightarrow \mathit{MEb} \underset{r_1}{\rightarrow} \mathit{1Mb},\\
\alpha_4 &= \top \Rightarrow \mathit{NMEb} \Rightarrow \mathit{MEb} \underset{r_1}{\rightarrow} \mathit{1Mb} \Rightarrow {}^1\!/_2Mb,\\
\alpha_5 &= \top \Rightarrow \mathit{NMEb} \underset{r_2}{\rightarrow} \neg\, {}^1\!/_2Mb,\\
\alpha_6 &= \top \underset{r_3}{\rightarrow} {}^1\!/_2Mb.
\end{aligned}
$$

[4] For support of this default, see Sloan (1997).

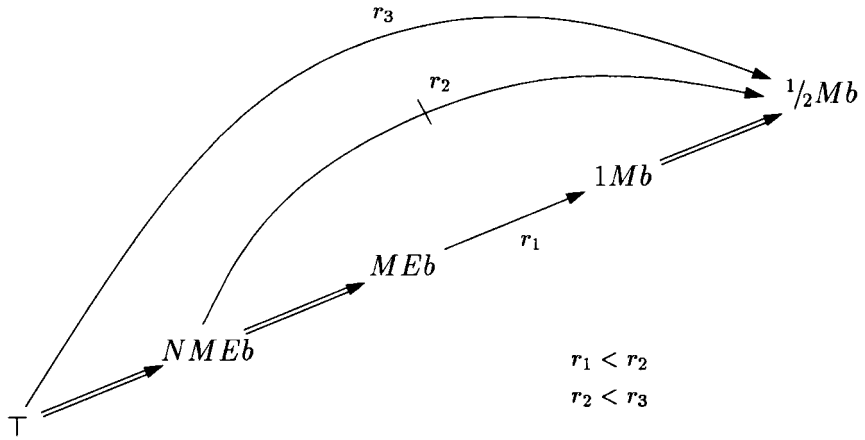

Figure 7. Is Beth a millionaire?

And given the ordering on defaults, the defeat relations are as follows: first, α_5 strictly defeats both α_4 and α_3; second, α_6 strictly defeats α_5. (The reason that α_5 defeats α_3 as well as α_4 is that α_4 is a strict extension of α_3; both α_3 and α_4 therefore support the statement $^1/_2Mb$ through the default rule r_1, which is weaker than the rule r_2 through which α_5 supports the conflicting statement $\neg\, ^1/_2Mb$.) Because α_6 strictly defeats α_5, it reinstates both α_3 and α_4, and so we should expect these two arguments to be endorsed by the PS logic. The iterative procedure again confirms this expectation, telling us that $F_\Gamma^1 = \{\alpha_1, \alpha_2, \alpha_6\}$, that $F_\Gamma^2 = \{\alpha_1, \alpha_2, \alpha_6, \alpha_3, \alpha_4\}$, and that $F_\Gamma^3 = F_\Gamma^2$, so that this set contains the justified arguments.

Looking only at the terminal conclusions of these various arguments, α_4 is no more problematic than what we have already seen in the wild chicken example: it is, at worst, a flawed argument for the correct conclusion $^1/_2Mb$, which is legitimately supported by the argument α_6. But the classification of α_3 among the justified arguments presents a more serious difficulty: the conclusion *1Mb* seems simply to be mistaken. Our only reason for believing that Beth has a million dollars is that she is a Microsoft employee, but this is overridden by the consideration that Beth is a new Microsoft employee, and is therefore unlikely to have even half a million dollars. As it happens, we do have an independent reason for believing that Beth has half a million dollars – but this gives us no reason at all to conclude that she has a million dollars. Reinstatement leads us to an incorrect conclusion.

4. The KT Logic

4.1. THE BACKGROUND FRAMEWORK

The KT logic, developed by Kowalski and Toni, is cast against the background of a framework developed earlier by these authors along with Bondarenko and Dung (1997). Although this framework is very general, and is shown to subsume a

number of familiar formalisms for nonmonotonic reasoning, we concentrate here on its application to the semantics of logic programs containing classical negation as well as negation as failure.

To fix notation, let us take $\mathcal{H}$ as the set of ground atoms from the language and $\mathcal{H}^+$ as $\mathcal{H}$ together with those formulas of the form $\neg A$ for A belonging to $\mathcal{H}$. The set *Literals* then includes the members of $\mathcal{H}^+$ as well as those formulas of the form $\sim A$ for A belonging to $\mathcal{H}^+$. The statement $\neg A$ represents the classical negation of A; the statement $\sim A$ represents the fact that A is not provable, or negation as failure.

An *extended logic program* is a set of rules of the form

$$A \Leftarrow B_1, \ldots, B_n,$$

with $n \geq 0$, with $B_1, \ldots, B_n$ belonging to *Literals*, and with A belonging to $\mathcal{H}^+$. Such a rule tells us that A can be inferred on the basis of $B_1, \ldots, B_n$. If we identify axioms with rules in which $n = 0$ – rules containing a conclusion but no premises – it is then straightforward to adapt the standard notion of an axiomatic proof to extended logic programs, resulting also in a standard concept $\vdash$ of derivability. Note that these rules do not permit backwards reasoning akin to modus tollens: in the rule displayed above, nothing in particular can be inferred from a negation of A. Note also that statements of the form $\sim A$ are permitted only among the premises of a rule, not in its conclusion; the language does not contain rules allowing us to infer that a statement is not provable.

The basic idea underlying the framework set out in Bondarenko et al. (1997) is that default reasoning should be seen as resulting from the supplementation of a theory with some suitable set of assumptions; such reasoning is nonmonotonic because adding statements to a theory might affect the suitability of certain assumptions, leading to the withdrawal of conclusions. In the case of theories expressed as a logic programs, the possible assumptions can be restricted to claims that certain formulas are not provable; more exactly, any possible *assumption* is a formula of the form $\sim A$.

Within this general framework, the notion of a suitable set of assumption can be characterized in a number of different ways, leading to different styles of default reasoning; but Kowalski and Toni favor a characterization that is based on the notion of admissibility. Fixing a logic program $\mathcal{P}$ as background, let us say that an assumption set Δ *attacks* a particular assumption $\sim A$ whenever $\mathcal{P} \cup \Delta \vdash A$; and let us say that one assumption set Δ *attacks* another assumption set Δ' whenever Δ attacks some assumption belonging to Δ'. The notion of an admissible set of assumptions can then be defined as follows:

- An assumption set Δ is *admissible* just in case (i) Δ does not attack itself, and (ii) if an assumption set Δ' attacks Δ, then Δ also attacks Δ'.

The first clause in this definition is simply a coherence condition; the intuition behind the second clause is that an assumption set is admissible if it is capable of "defending" itself, by attacking any other assumption set that attacks it.

Once the suitable assumption sets are identified with the admissible sets, it is a simple matter to define a notion of credulous consequence for logic programs:

- A statement A is a *credulous consequence* of a logic program $\mathcal{P}$ just in case $\mathcal{P} \cup \Delta \vdash \mathcal{A}$ for some admissible set Δ.

It may then seem natural to characterize the skeptical consequences of a logic program as those that follow whenever that program is supplemented with any admissible set, but this characterization turns out to be overly restrictive since a number of relatively uninformative assumption sets, including the empty set, are admissible. It is more reasonable first to introduce the notion of a *preferred* assumption set, defined as a maximal admissible set, and then to rely on this notion in the characterization of skeptical consequence:

- A statement A is a *skeptical consequence* of a logic program $\mathcal{P}$ just in case $\mathcal{P} \cup \Delta \vdash \mathcal{A}$ for each preferred assumption set Δ.

4.2. DEFEASIBLE RULES

The approach as presented so far, which is entirely contained in Bondarenko et al. (1997), applies only to extended logic programs as defined above, involving only strict rules. What Kowalski and Toni (1996) add is a way of generalizing this approach to apply also to a class of *defeasible* extended logic programs, which might contain, in addition to these strict rules, also certain defeasible rules of the form

$$r: \quad A \leftarrow B_1, \ldots, B_n, \tag{1}$$

where r functions as a label of the rule, and where A and $B_1, \ldots, B_n$ are subject to restrictions identical to those set out above. Such a rule tells us that the statements $B_1, \ldots, B_n$ provide a reason for drawing the conclusion A, even if the reason is not conclusive.

Among the rules of a defeasible logic program, there may be certain rules, themselves either strict or defeasible, whose purpose is to specify priority rankings among various defeasible rules, referring to these rules through their labels. As an example, the rule

$$r < r' \Leftarrow,$$

would tell us that the defeasible rule r' is always to be given greater priority than the defeasible rule r.

There is an evident similarity between the defeasible extended logic programs described here and the earlier ordered theories of Prakken and Sartor; it is a simple matter to translate information presented in one formalism into the other. In analyzing this information, however, Kowalski and Toni follow a different strategy from that of Prakken and Sartor. Rather than presenting a semantic account that applies directly to logic programs containing defeasible rules, Kowalski and Toni instead

suggest a general scheme for transforming any defeasible extended logic program into an ordinary extended logic program, containing no defeasible rules at all, but containing a few additional predicates: *holds*, *defeated*, and *conflict*. The ordinary logic program into which the defeasible logic program is transformed is supposed to provide an accurate and precise representation of the meaning of the original defeasible program.

The idea behind the transformation scheme is simple. Each defeasible rule of the form (1) above is to be replaced with two strict rules of the form

$$A \Leftarrow holds(r),$$
$$holds(r) \Leftarrow B_1, \ldots, B_n, \sim defeated(r),$$

telling us, intuitively, that the consequent A of the original rule can be established if the rule "holds," and that the rule holds if its antecedent formulas $B_1, \ldots, B_n$ can be established and the rule itself is not shown to be defeated.[5] The notion of defeat is then characterized through the single rule

$$defeated(r) \Leftarrow r < r', conflict(r, r'), holds(r'), \tag{2}$$

telling us that the rule r is defeated whenever a rule r' that has a higher priority than r and that also conflicts with r can be shown to hold.

Unfortunately, Kowalski and Toni do not provide a full definition of the notion of conflict that figures in their characterization of defeat; the idea is simply illustrated with rules such as

$$conflict(r, r') \Leftarrow conclusion(r, A), conclusion(r', \neg A),$$

where $conclusion(r, A)$ is supposed to hold whenever A is the conclusion of the rule r. This partial definition is sufficient for explicit conflicts, between rules with complementary conclusions; but it is not able to capture the kind of implicit conflicts illustrated earlier in Figure 3, where, as we recall, the rule r_1 supporting A and the rule r_2 supporting B conflict in the presence of the strict information telling us that A implies C and that B implies $\neg C$. In order to complete the definition of their transformation scheme, Kowalski and Toni would have to provide a treatment that handles implicit as well as explicit conflicts.

For present purposes, however, we avoid the issue of actually characterizing the notion of conflict, and use the *conflict* predicate instead simply to list the rules that a proper definition would classify as conflicting; our only general assumption is the notion of conflict is symmetric, expressed through the rule

$$conflict(r, r') \Leftarrow conflict(r', r). \tag{3}$$

[5] Kowalski and Toni note that a rule label r is used ambiguously in the transformed program, to refer both to the original defeasible rule and to its conclusion.

4.3. EXAMPLES

In order to illustrate the transformation scheme, let us turn first to the Tweety Triangle from Figure 1. The information from this diagram can be presented as a defeasible extended logic program containing the two strict rules

$$\begin{array}{l} Pt \Leftarrow, \\ Bt \Leftarrow Pt, \end{array} \tag{4}$$

telling us that Tweety is a penguin, and that Tweety is a bird if he is a Penguin; the two defeasible rules

$$\begin{array}{l} r_1 : \ Ft \leftarrow Bt, \\ r_2 : \ \neg Ft \leftarrow Pt, \end{array}$$

providing us with a reason to believe that Tweety flies, since he is a bird, and also with a reason for believing Tweety does not fly, since he is a penguin; and the priority ranking

$$r_1 < r_2 \Leftarrow, \tag{5}$$

telling us that the second of these reasons is stronger than the first. Following the scheme, this defeasible program can be transformed into an ordinary program by replacing the first of its defeasible rule with the pair of strict rules

$$\begin{array}{l} Ft \Leftarrow \mathit{holds}(r_1), \\ \mathit{holds}(r_1) \Leftarrow Bt, \sim \mathit{defeated}(r_1), \end{array} \tag{6}$$

by replacing the second of its defeasible rules with the strict rules

$$\begin{array}{l} \neg Ft \Leftarrow \mathit{holds}(r_2), \\ \mathit{holds}(r_1) \Leftarrow Pt, \sim \mathit{defeated}(r_2), \end{array} \tag{7}$$

and by supplementing the program with the statement

$$\mathit{conflict}(r_1, r_2) \Leftarrow \tag{8}$$

to capture the appropriate conflict relations. The transformed program $\mathcal{P}$ representing the original information thus includes the rules (4), (5), (6), (7), and (8), as well as the background rules (2) and (3) concerning the notions of defeat and conflict.

In calculating the admissible assumption sets for this logic program, we can restrict our attention to assumptions of the form $\sim \mathit{defeated}(r)$, since these are the only assumptions that the program contains. Apart from the empty set, then, there are only three assumption sets to consider:

$$\begin{array}{l} \Delta_1 = \ \{\sim \mathit{defeated}(r_1)\}, \\ \Delta_2 = \ \{\sim \mathit{defeated}(r_2)\}, \\ \Delta_3 = \ \{\sim \mathit{defeated}(r_1), \sim \mathit{defeated}(r_2)\}. \end{array}$$

When added to the program $\mathcal{P}$, the information from the assumption set Δ_3 immediately yields the contradictory conclusions Ft and $\neg Ft$, and so both *defeated*(r_1) and *defeated*(r_2). The set Δ_3 therefore attacks itself, and so violates the first clause of the admissibility definition. Turning to Δ_1, we can see that this set is attacked by Δ_2, since Δ_2 together with $\mathcal{P}$ allows us to derive *defeated*(r_1). But Δ_1 does not itself attack Δ_2: although Δ_1 together with $\mathcal{P}$ allows us to derive *holds*(r_1), and we know *conflict*(r_1, r_2), we do not have $r_2 < r_1$ as required by the rule (2) for a justification of *defeated*(r_2). Since Δ_2 attacks Δ_1, but Δ_1 cannot defend itself by attacking Δ_2, it violates the second clause of the admissibility definition.

Only Δ_2 is an admissible assumption set: it does not attack itself, and it defends itself against the only set Δ_3 that attacks it. The set Δ_2 is, moreover, the unique preferred assumption set, since the only other admissible assumption set is the empty set, and Δ_3 is maximal among these two. When supplemented with the information from this unique preferred assumption set, $\mathcal{P}$ yields $\neg Ft$ as a conclusion. This statement is therefore a skeptical consequence of this program, as desired.

The treatment of the Nixon Diamond from Figure 2 is in many ways similar. The defeasible logic program representing this information contains the two strict rules

$$\begin{array}{l} Qn \Leftarrow, \\ Rn \Leftarrow, \end{array} \tag{9}$$

telling us that Nixon is a Quaker and a Republican, as well as the rules

$$\begin{array}{l} r_1 : \ Pn \leftarrow Qn, \\ r_2 : \ \neg Pn \leftarrow Rn, \end{array}$$

providing defeasible reasons to believe that Nixon is a pacifist if he is Quaker, and that he is not a pacifist if he is a Republican. In carrying out the transformation, these two defeasible rules are replaced, as before, by the four strict rules

$$\begin{array}{l} Pn \Leftarrow holds(r_1), \\ holds(r_1) \Leftarrow Qn, \sim defeated(r_1), \\ \neg Pn \Leftarrow holds(r_2), \\ holds(r_2) \Leftarrow Rn, \sim defeated(r_2), \end{array} \tag{10}$$

and then supplemented with the statement

$$conflict(r_1, r_2) \Leftarrow \tag{11}$$

reflecting the conflict between the defeasible rules.

In our defeasible representation of the Tweety Triangle, the rule concerning penguins was explicitly assigned a higher priority than the rule concerning birds through the statement (5), which was carried over into the strict program. In the present case, by contrast, neither of the rules r_1 nor r_2 is explicitly given a higher

priority than the other in the defeasible representation, and so it is natural to think that the strict transformation of this program would likewise fail to register any priority between these rules. Kowalski and Toni take a different approach, however: "where priorities are not given explicitly, ... we treat each rule as having a higher priority than every other rule with a contradictory conclusion" (Kowalski and Toni 1996). Following these instructions, the strict transformation of the Nixon diamond should be supplemented also with the statements

$$\begin{array}{l} r_1 < r_2 \Leftarrow, \\ r_2 < r_1 \Leftarrow, \end{array} \tag{12}$$

giving each of these rules a higher priority than the other. The strict program $\mathcal{P}$ representing the Nixon Diamond thus contains the rules (9), (10), (11), and (12), as well as the background rules (2) and (3).

It is easy to see that both $\Delta_1 = \{\sim \mathit{defeated}(r_1)\}$ and $\Delta_2 = \{\sim \mathit{defeated}(r_2)\}$ are preferred assumption sets for this program. When supplemented with Δ_1, the program $\mathcal{P}$ yields Pn but not $\neg Pn$ as a conclusion; when supplemented with Δ_2, the program yields $\neg Pn$ but not Pn as a conclusion. Since the program does not yield either Pn or $\neg Pn$ when supplemented with every preferred assumption set, the skeptical interpretation of the KT logic does not allow us to conclude either that Nixon is a pacifist or that he is not, as desired.

5. Problems with the KT Logic

Like the PS logic, then, the KT logic yields the desired results when applied to the Tweety Triangle and the Nixon Diamond, two familiar benchmark examples; and Kowalski and Toni show that the theory gives an adequate treatment of several richer and more complex reasoning scenarios.

As it turns out, however, the KT logic also exhibits two problems closely analogous to those faced by the PS logic – concerning argument construction and evaluation, and also reinstatement – which can be illustrated by the same examples set out earlier.

5.1. ARGUMENT CONSTRUCTION AND EVALUATION

In order to illustrate the first problem we formulate the example from Figure 4 in the KT logic. The representation of this information as a defeasible logic program would contain the strict rules

$$\begin{array}{l} PDa \Leftarrow, \\ Ra \Leftarrow, \\ La \Leftarrow PDa, \\ Ba \Leftarrow Ra, \end{array} \tag{13}$$

telling us that Ann is a public defender and a renter in Brentwood, that she is a lawyer if she is a public defender, and that she is a resident of Brentwood if she rents in Brentwood; the defeasible rules

$$\begin{aligned} &r_1 : \ Wa \leftarrow La, \\ &r_2 : \ \neg Wa \leftarrow PDa, \\ &r_3 : \ Wa \leftarrow Ba, \\ &r_4 : \ \neg Wa \leftarrow Ra, \end{aligned}$$

giving us reason to believe that Ann is wealthy if she is a lawyer but not wealthy if she is a public defender, and that Ann is wealthy if she is a resident of Brentwood but not wealthy if she is a renter in Brentwood; and the priority information

$$\begin{aligned} &r_1 < r_2 \Leftarrow, \\ &r_3 < r_4 \Leftarrow \end{aligned} \tag{14}$$

telling us that the defeasible rule concerning public defenders is to be given greater weight than that concerning lawyers, and that the defeasible rule concerning Brentwood renters is to be given greater weight than that concerning Brentwood residents.

In transforming this information into a strict logic program, the four defeasible rules are replaced by the eight strict rules

$$\begin{aligned} &Wa \Leftarrow holds(r_1), \\ &holds(r_1) \Leftarrow La, \sim defeated(r_1), \\ &\neg Wa \Leftarrow holds(r_2), \\ &holds(r_2) \Leftarrow PDa, \sim defeated(r_2), \\ &Wa \Leftarrow holds(r_3), \\ &holds(r_1) \Leftarrow Ba, \sim defeated(r_3), \\ &\neg Wa \Leftarrow holds(r_4), \\ &holds(r_4) \Leftarrow Ra, \sim defeated(r_4), \end{aligned} \tag{15}$$

and the conflicts among these various rules is registered:

$$\begin{aligned} &conflict(r_1, r_2) \Leftarrow, \\ &conflict(r_1, r_4) \Leftarrow, \\ &conflict(r_3, r_2) \Leftarrow, \\ &conflict(r_3, r_4) \Leftarrow . \end{aligned} \tag{16}$$

As far as priorities, the explicitly provided information from (14) is carried over to the strict program; but in addition, following the recipe provided by Kowalski and Toni, conflicting rules for which no explicit priorities are provided must each must each be given a higher priority than the other. No priorities are provided for the

conflicting rules r_1 and r_4, or for the conflicting rules r_3 and r_3, and so the program must be supplemented with the statements

$$\begin{array}{l} r_1 < r_4 \Leftarrow, \\ r_4 < r_1 \Leftarrow, \\ r_2 < r_3 \Leftarrow, \\ r_3 < r_2 \Leftarrow . \end{array} \tag{17}$$

The overall strict representation of the information from Figure 4 is thus the program $\mathcal{P}$ containing (13), (14), (15), (16), and (17), as well as (2) and (3).

This strict program allows for two preferred assumption sets:

$$\Delta_1 = \{\sim \mathit{defeated}(r_1), \sim \mathit{defeated}(r_3)\},$$
$$\Delta_2 = \{\sim \mathit{defeated}(r_2), \sim \mathit{defeated}(r_4)\}.$$

The second of these is intuitively attractive, and yields the results supported by skeptical inheritance reasoning: when supplemented with the assumption set Δ_2, the program $\mathcal{P}$ yields the conclusion $\neg Wa$. The first assumption set, however, seems less attractive: when supplemented with Δ_1, the program $\mathcal{P}$ yields the conclusion Wa. Since $\neg Wa$ does not follow from both preferred assumption sets, this statement is not generated as a skeptical conclusion. Contrary to the intuitions underlying inheritance reasoning, we are not able to conclude that Ann is not wealthy.

What is the difficulty? From an intuitive point of view, it seems that the assumption set Δ_1 should be, in some sense, overridden or preempted by the set Δ_2: the rule r_1 is conclusively defeated by the rule r_2 and the rule r_3 is conclusively defeated by the rule r_4. In fact, Δ_2 does attack Δ_1, since r_2 defeats r_1 and r_4 defeats r_3. But according to the KT logic, it turns out that Δ_1 is able to defend itself from this attack, since r_3 defeats r_2 and r_1 defeats r_4. The problem here is that neither assumption belonging to this set – neither $\sim \mathit{defeated}(r_1)$ nor $\sim \mathit{defeated}(r_3)$ – should be tenable. When taken as a pair, however, each of these untenable assumptions is able to buttress the other by attacking the assumption that should conclusively defeat it, like two dissemblers each defending the veracity of the other.

Although the KT logic does not involve reasoning with explicitly constructed arguments, it is nevertheless possible to see the current difficulty presented by Figure 4 as analogous to the problem that this example presented earlier, for the PS logic. In the earlier case, the example was used to show that the idea of constructing all arguments prior to the process of evaluation allowed certain unacceptable arguments – α_5 and α_7 from that discussion – to perturb the overall pattern of defeat. An iterative approach, similar to that followed in the inheritance literature, would avoid this problem by requiring arguments to be evaluated upon construction, and then immediately discarded if found to be unacceptable.

In the present case, the policy of considering all possible assumption sets is in many ways like the previous policy of considering all possible arguments. The

assumptions $\sim$ *defeated*(r_1) and $\sim$ *defeated*(r_3) are like the earlier arguments α_5 and α_7, supporting the conclusion that Ann is wealthy because she is a lawyer or a Brentwood resident. These assumptions should not enter into consideration, but should be individually defeated by the assumptions $\sim$ *defeated*(r_2) and $\sim$ *defeated*(r_4), supporting the conclusion that Ann is not wealthy because she is a public defender and a Brentwood renter. Again, it may be worth exploring an iterative approach, according to which assumptions are ordered in accord with the arguments they support, and then evaluated – and either accepted or rejected – in that order.

5.2. REINSTATEMENT

We illustrate the second problem by providing a KT logic formulation of the example from Figure 7. The representation of this information as a defeasible logic program would include the strict rules

$$\begin{array}{l} NMEb \Leftarrow, \\ MEb \Leftarrow NMEb, \\ {}^{1}\!/_{2}Mb \Leftarrow 1Mb, \end{array} \tag{18}$$

telling us that Beth is a new Microsoft employee, that she is a Microsoft employee if she is a new Microsoft employee, and that she has half a million dollars if she has a million dollars; the defeasible rules

$$\begin{array}{l} r_1 : \ 1Mb \leftarrow MEb, \\ r_2 : \ \neg\, {}^{1}\!/_{2}Mb \leftarrow NMEb, \\ r_3 : \ {}^{1}\!/_{2}Mb \leftarrow \ , \end{array}$$

giving us reason to believe that Beth has a million dollars if she is a Microsoft employee, that she does not have even half a million dollars if she is a new Microsoft employee, but that she in fact has half a million dollars; and the priority information

$$\begin{array}{l} r_1 < r_2 \Leftarrow, \\ r_2 < r_3 \Leftarrow, \end{array} \tag{19}$$

telling us that the defeasible information concerning new Microsoft employees is to be given greater weight than the information concerning Microsoft employees in general, but that the defeasible information concerning Beth in particular is to be given even greater weight than that concerning new Microsoft employees.

Transforming this representation into a defeasible program, we replace the three defeasible rules listed above by the six strict rules

$$\begin{array}{l} \mathit{1Mb} \Leftarrow \mathit{holds}(r_1), \\ \mathit{holds}(r_1) \Leftarrow \mathit{MEb}, \sim \mathit{defeated}(r_1), \\ \neg\, {}^{1}\!/_{2}\mathit{Mb} \Leftarrow \mathit{holds}(r_2), \\ \mathit{holds}(r_2) \Leftarrow \mathit{NMEb}, \sim \mathit{defeated}(r_2), \\ {}^{1}\!/_{2}\mathit{Mb} \Leftarrow \mathit{holds}(r_3), \\ \mathit{holds}(r_1) \Leftarrow \sim \mathit{defeated}(r_3), \end{array} \tag{20}$$

and supplement the result with statements registering the conflict between the rule r_2 and the rules r_1 and r_3:

$$\begin{array}{l} \mathit{conflict}(r_1, r_2) \Leftarrow, \\ \mathit{conflict}(r_2, r_3) \Leftarrow . \end{array} \tag{21}$$

Since explicit priorities are listed for each pair of conflicting rules, no new priorities must be added, and so the strict representation of the information from this example is the program $\mathcal{P}$ containing (18), (19), (20), and (21), along with (2) and (3).

It is easy to see that this strict program allows for exactly one preferred assumption set:

$$\Delta_1 = \{\sim \mathit{defeated}(r_1), \sim \mathit{defeated}(r_3)\}.$$

When taken together with the program $\mathcal{P}$, the assumption set Δ_1 yields both ${}^{1}\!/_{2}\mathit{Mb}$ and *1Mb* as conclusions. The skeptical interpretation of the KT logic thus allows us to conclude, correctly, that Beth has half a million dollars, but also – contrary to the intuitions underlying skeptical inheritance reasoning – that Beth has a million dollars.

Again, although the KT logic does not involve the consideration of explicitly constructed arguments, it is possible to see the current difficulty presented by Figure 7 as analogous to the difficulty that this example presented for the PS logic – a matter of reinstatement. Intuitively, it seems that the assumption $\sim \mathit{defeated}(r_1)$, supporting the conclusion that Beth has a million dollars since she is a Microsoft employee, should be defeated by the assumption $\sim \mathit{defeated}(r_2)$, telling us that she does not have even half a million dollars since she is only a new Microsoft employee. Why, then, is Δ_1 not undermined by the assumption set $\Delta_2 = \{\sim \mathit{defeated}(r_2)\}$? In fact, Δ_2 does attack Δ_1. But Δ_1 is able to defend itself against this attack, since it also contains the assumption $\sim \mathit{defeated}(r_3)$, telling us that Beth does have half a million dollars, and so attacking the assumption $\sim \mathit{defeated}(r_2)$. In allowing Δ_1 to defend itself against the attack from Δ_2, the assumption $\sim \mathit{defeated}(r_3)$ therefore attacks the assumption that attacks the assumption $\sim \mathit{defeated}(r_1)$, and so, according to the theory, reinstates $\sim \mathit{defeated}(r_1)$ as a legitimate assumption. It is this assumption that supports the result that Beth has a million dollars: reinstatement again leads to a peculiar conclusion.

6. Conclusion

Argument systems offer a promising approach both in the analysis of defeasible reasoning and also, more recently, as a foundation for negotiation and dispute-resolution protocols in multi-agent systems. Nevertheless, these formalisms are often complicated and occasionally based on unclear principles.

In this paper I have argued that two of the cleanest and most carefully developed recent argument systems – the PS logic and the KT logic – support incorrect conclusions when applied in the simple domain of defeasible inheritance networks. Although it is always difficult, in evaluating a system with multiple, interacting components, to attribute problems to any one particular feature, I have identified two ideas in these argument systems, prominent also in a number of other argument-based formalisms for defeasible reasoning, that seem to be responsible for the difficulties. The first is the idea that the set of acceptable arguments should be determined on the basis of a pattern of defeat relations holding among all possible arguments, including those arguments that are to be definitively rejected. The second is the idea of reinstatement – that an argument is acceptable, even if defeated, as long as all the arguments that defeat it are themselves defeated. The examples set out in this paper suggest that both of these ideas are mistaken.

Acknowledgements

I have received useful comments on this paper from Bob Kowalski and Henry Prakken, both of whom have suggested various ways in which their systems can be modified to avoid the problems described here.

References

Bondarenko, A., Dung, P. M., Kowalski, R. and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. Artificial Intelligence 93: 63–101.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reason, logic programming, and n-person games. Artificial Intelligence 77: 321–357.

Dung, P. M., Kowalski, R. and Toni, F. (1996). Synthesis of proof procedures for default reasoning. In Proceedings of the International Workshop on Logic Program Synthesis and Transformation (LOPSTR'96), Springer Verlag Lecture Notes in Computer Science # 1207, 313–324. Springer Verlag.

Etherington, D. (1987). Formalizing nonmonotonic reasoning systems. Artificial Intelligence 31: 41–85.

Etherington, D. and Reiter, R. (1983). On inheritance hierarchies with exceptions. In Proceedings of the Third National Conference on Artificial Intelligence (AAAI-83), 104–108.

Horty, J. (1994). Some direct theories of nonmonotonic inheritance. In Gabbay, D., Hogger, C. and Robinson, J. (eds.), Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning, 111–187. Oxford University Press.

Horty, J. and Thomason, R. (1988). Mixing strict and defeasible inheritance. In Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88), 427–432.

Horty, J., Thomason, R. and Touretzky, D. (1990). A skeptical theory of inheritance in nonmonotonic semantic networks. Artificial Intelligence 42: 311–348.

Kowalski, R. and Toni, F. (1996). Abstract argumentation. Artificial Intelligence and Law 4: 275–296.

Pollock, J. (1987). Defeasible reasoning. Cognitive Science 11: 481–518.

Prakken, H. and Sartor, G. (1996). A dialectical model of assessing conflicting arguments in legal reasoning. Artificial Intelligence and Law 4: 331–368.

Prakken, H. and Sartor, G. (1996). A system for defeasible argumentation with defeasible priorities. In Proceedings of the International Conference on Formal Aspects of Practical Reasoning, Springer Verlag Lecture Notes in Artificial Intelligence # 1085. Springer Verlag.

Prakken, H. and Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities. Journal of Applied Non-classical Logics 7: 25–75.

Prakken, H. and Sartor, G. (1998). Modelling reasoning with precedents in a formal dialogue game. Artificial Intelligence and Law 6: 231–287.

Prakken, H. and Vreeswijk, G. (forthcoming). Logics for defeasible argumentation. In Gabbay, D. (ed.), Handbook of Philosophical Logic (Second Edition). Kluwer Academic Publishers.

Reiter, R. and Criscuolo, G. (1981). On interacting defaults. In Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI-81), 270–276.

Sloan, A. (1997). Millionaires next door – that is, if your neighbors work for Microsoft. Newsweek, December 8.

Touretzky, D., Horty, J. and Thomason, R. (1987). A clash of intuitions: the current state of nonmonotonic multiple inheritance systems. In Proceedings of the Tenth International Joint Conference on Artificial Intelligence(IJCAI-87), 476–482. Morgan Kaufmann.

Touretzky, D., Thomason, R. and Horty, J. (1991). A skeptic's menagerie: conflictors, preemptors, reinstaters, and zombies in nonmonotonic inheritance. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91), 478–483. Morgan Kaufmann Publishers.